COMMUNICATING QUALITY TO MAKE THE BEST USE OF HETEROGENEOUS DATA

Lucy Bastin
Aston University
and
Joint Research Centre of the European Commission

MY AREAS OF INTEREST

Quantification and communication of uncertainty and quality

- Allowing appropriate handling in web-based models & workflows
- Enriching the information which data publishers record
- Capturing user feedback about data use and issues discovered in practice

Documenting the nature and quality of citizen science data

- Mobilising CS data so it can be discovered, aggregated and reused

Digital Observatory for Protected Areas

Delivering data on biodiversity values and threats to support decision making.

- eg. Species irreplaceability, habitat diversity, ecoregion protection…
- eg. Pressure from agriculture, roads, resource extraction…

INTAMAP – WEB-BASED INTERPOLATION

ELSEVIER

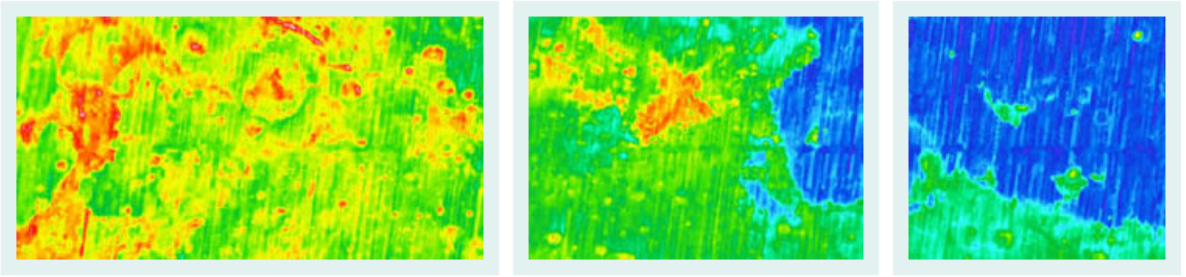
Computers & Geosciences

Volume 37, Issue 3, March 2011, Pages 343-352

INTAMAP: The design and implementation of an interoperable automated interpolation web service

INTAMAP is a Web Processing Service for the automatic spatial interpolation of measured point data. Requirements were (i) using open standards for spatial data such as developed in the context of the Open Geospatial Consortium (OGC), (ii) using a suitable environment for statistical modelling and computation, and (iii) producing an integrated, open source solution. The system couples an open-source Web Processing Service (developed by 52°North), accepting data in the form of standardised XML documents (conforming to the OGC Observations and Measurements standard) with a computing back-end realised in the R statistical environment. The probability distribution of interpolation errors is encoded with UncertML, a markup language designed to encode uncertain data. Automatic interpolation needs to be useful for a wide range of applications and the algorithms have been designed to cope with anisotropy, extreme values, and data with known error distributions. Besides a fully automatic mode, the system can be used with different levels of user control over the interpolation process.

UncertML describing and exchanging uncertainty...



| About UncertML | API documentation | User guide

URI:	http://www.uncertml.org/distributions/normal
UncertML name:	NormalDistribution
Alternative names:	Gaussian distribution
Definition:	A random variable x is normally distributed if the probability density function (pdf) is of the form shown below. The distribution is usually denoted as $x \sim \mathcal{N}(\mu, \sigma^2)$ where μ is known as the mean parameter and σ^2 the variance parameter. If the random variable x is a vector of length greater than one, the normal distribution can be generalised to the Multivariate normal. A reason for the widespread usage

UncertML – a schema & set of terms for communicating probabilistic aspects of data quality.

Designed to be used in combination with other models and standards: e.g., PROV-O for provenance, ISO 19115 / 19157, O&M …

```
<gmd:DQ_QuantitativeAttributeAccuracy>
   <gmd:result>
      <gmd:DQ_QuantitativeResult>
         <gmd:valueType>
         <gco:RecordType xlink:href="http://www.uncertml.org/distributions/normal">
                  Value of the vertical DEM accuracy
            </gco:RecordType>
         </gmd:valueType>
         <gmd:valueUnit>m</gmd:valueUnit>
      <gmd:value>
      <gco:Record>
                  <un:NormalDistribution>
                        <un:mean>1.2</un:mean>
                        <un:variance>3.6</un:variance>
                  </un:NormalDistribution>
               </gco:Record>
         </gmd:value>
      </gmd:DQ_QuantitativeResult>
   </gmd:result>
</gmd:DQ_QuantitativeAttributeAccuracy>
```

https://github.com/uncertweb/uncertml/blob/master/UncertML30.xsd

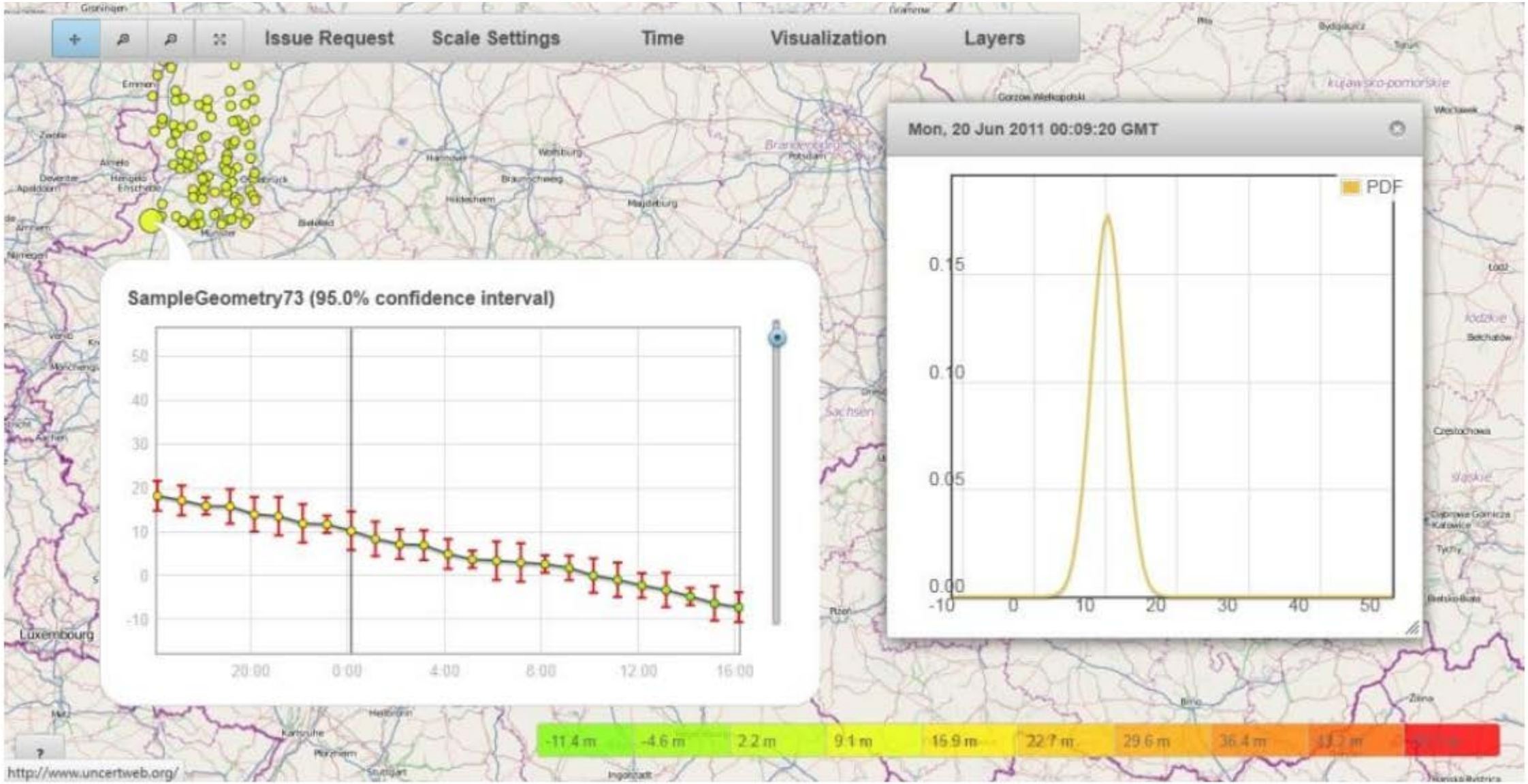
UNCERTWEB

- Considered the ModelWeb, where data and models are designed to be discovered and orchestrated in workflows.

- Uncertainty is rarely documented in the data or the models, but the reliability of the results MUST be evaluated and communicated to decision-makers.

- Extended the development of UncertML, with APIs (R and Java), schema encodings, and tools for expert elicitation, model emulation, visualisation of uncertainty

Open Layers based client for the visualization of uncertain observations.



https://wiki.52north.org/bin/view/Geostatistics/Greenland

GEOVIQUA

1) Allow producers to provide enriched metadata (especially on provenance).

2) Allow users to supply data reviews & quality reports, post hoc.

Allowing metadata to be continuously enriched can help in assessing **fitness-for-purpose** – for different users, with different needs and quality criteria
http://www.opengeospatial.org/standards/guf

Open annotations, commentary metadata…

Producer metadata and user feedback are combined to produce a label which tells you how well documented the dataset is.

http://wiki.ieee-earth.org/Documents/GEOSS_Tutorials/GEOSS_Provider_Tutorials/Data_Quality_Tutorial_for_GEOSS_Providers

4.1 Use cases: producer quality model

4.1.1 Citing a publication within a metadata document

4.1.2 Producer supplies 'soft knowledge' or advice on issues discovered with the dataset which cannot be easily encoded elsewhere

4.1.3 Recording the traceability of a quality statement

4.1.4 Citing one or more datasets used as reference for a quality evaluation

4.1.5 Providing full statistical information on the results of quality assessments

4.2 Use cases: user quality model

4.2.1 A rating based on a dataset as a whole

4.2.2 A rating based on a spatial or temporal subset of the data

4.2.3 A domain-specific rating in the context of weather forecasting

4.2.4 Justification for a rating

4.2.5 Provide general feedback on a complete dataset

4.2.6 Provide general feedback on a part of the data

4.2.7 Provide general feedback on a metadata record

4.2.8 Provide general feedback on a specific metadata record elements

4.2.9 Add a domain-specific comment to a data set

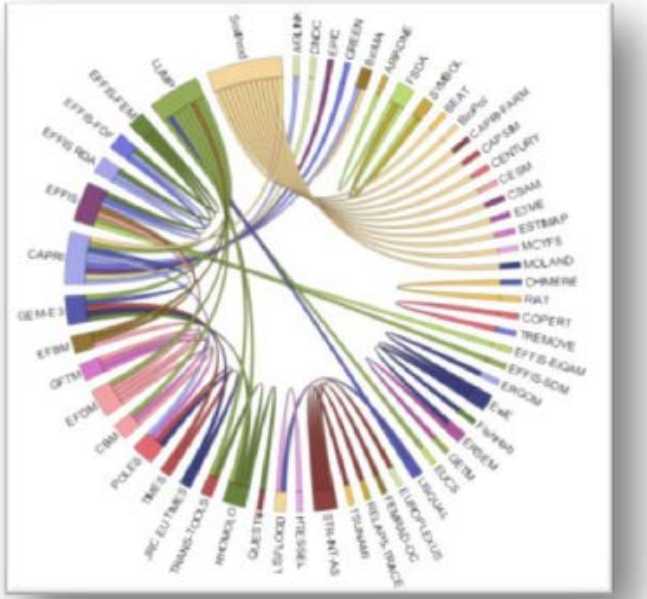
4.2.10 Start a thematic discussion on a dataset

4.2.11 Report external feedback related to a dataset

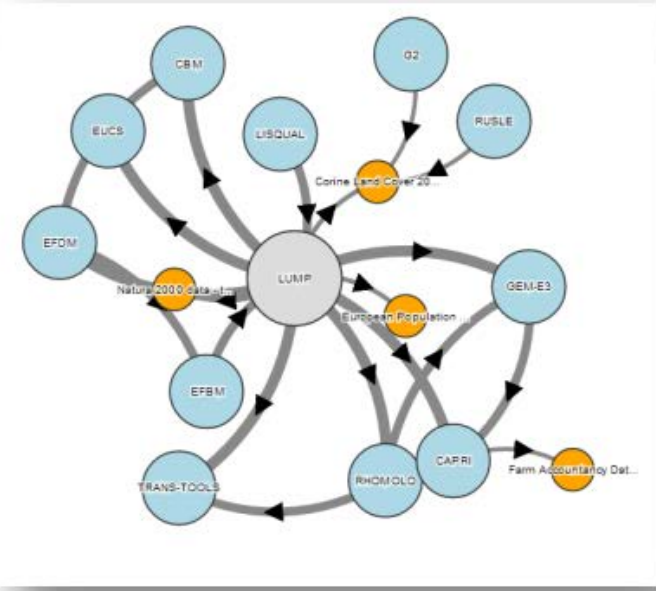
4.2.12 Report external feedback related to a part of a dataset

4.2.13 Report publication which cites a dataset

MIDAS makes links between models explicit.



Who uses what? MIDAS links models and data.



STANDARDS FOR CITIZEN SCIENCE

Citizen science is exploding, but much data is unavailable for reuse.

- Challenging to aggregate & make sense of these heterogeneous resources.
  - Units, phenomena, taxonomic definitions, precision, sampling protocol…
- Data quality information will come in many forms.
  - Reported, calculated *post hoc*, added by users, inferred from method / reputation…
  - Numerous groups are currently moving towards formal or informal solutions – now is a great time to coordinate.
  - COST action CA15212, TDWG (SamplingEvent in Extended Darwin Core), Atlas of Living Australia, OGC, ECSA and CSA, RDA/CODATA, W3C Data Quality vocabularies, Wilson Centre…

http://citizenscience.org/association/about/working-groups/data-and-metadata-working-group/

QUALITY DOCUMENTATION IN CITIZEN SCIENCE

It is proposed that the PPSR-CORE project element should point at a quality report at a resolvable URL.

Challenge: to encourage quality documentation of any sort, while ensuring that records can be interpreted and aggregated for re-use and future fitness-for-purpose assessment.

PRACTICAL USE CASE

Project – Invasive Alien Species app from JRC

Quality PROTOCOL: clear series of QA steps

Quality RESULT: information on, e.g., spatial precision of records

Proposal:

1) practically encode the above in a variety of ways so users can identify a strategy that fits with their current capacity.

2) document best practice for each style of documentation (e.g., selected controlled vocabularies)

3) work on a ‘worldview’ which aggregates and crosswalks the standards.

CANDIDATE ENCODINGS

1) Descriptive web page

2) Descriptive web page, annotated to link each step to accepted terms and clarify the sequence / relationships

3) ISO 19157 document with QA embedded in the Lineage*

4) ISO 19157 with Traceability statement for the Quality Statement*

5) 'UncertProv' a la Car, Cox & Fitch*

6) Any other user-friendly approach… suggestions welcome!!

* combines protocol and result)

Associating uncertainty with datasets using Linked Data and allowing propagation via provenance chains

Nicholas Car, Simon Cox, and Peter Fitch

UncertProv extends the PROV-O provenance ontology with an RDF formulation of the UncertML conceptual model elements, adds further elements to support uncertainty representation without a conceptual model and the integration of UncertML through links to documents. The Linked ID API provides a systematic way of navigating from dataset objects to their UncertProv metadata and back again. The Linked Data API's 'views' capability enables access to UncertML and non-UncertML uncertainty metadata representations for a dataset.

Thank you for your time!

lucy.bastin@ec.europa.eu